Examining the evolutionary arms race between transposons and genome defense provided by the piRNA pathway

By

Xi Chen

Submitted to the graduate degree program in Ecology and Evolutionary Biology and Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts

_____

Chairperson (Justin Blumenstiel, Ph.D.)

_____

(Lena Hileman, Ph.D.)

_____

(Maria Orive, Ph.D.)

Date Defended: September 4th 2015

The Thesis committee for Xi Chen certifies that this is the approved version of the following thesis:

Examining the evolutionary arms race between transposons and genome defense provided by the piRNA pathway

_____

Chairperson (Justin Blumenstiel, Ph.D.)

Date Approved: September 18th 2015

ABSTRACT

Transposable elements (TEs) are often referred to as selfish elements that cause harmful mutations by their mobilization in genome. TEs use a “copy and paste” or "cut and paste" mechanism that allows rapid exponential growth within populations, and these TE insertions result in many harmful mutations. Thus, scientists are interested in understanding what prevents TEs from unlimited proliferation. Natural selection has long been considered to be a major factor controlling spread of TEs. Recently, genome defense by the piRNA pathway has proved to be another key mechanism in controlling TEs’ proliferation. A recent study of the molecular evolution of the piRNA pathway indicated that several piRNA genes show a high rate of adaptive evolution rate (Obbard, Gordon et al. 2009). Thus, we predict that there probably is an evolutionary arms race between TEs and the genome defense provided by the piRNA machinery. Like host-parasite dynamics, co-evolution could be driven by adaptions and counter-adaptions in TEs and the piRNA pathway. However, a previous study has shown that the piRNA machinery in *Drosophila species* with higher TE content show greater levels of purifying selection (Castillo, Mell et al. 2011). Interestingly, they found that codon bias is increased in the piRNA genes of *Drosophila* species with higher TE abundance. These results may indicate that there is an evolutionary arms race between TEs and host defense provided by the piRNA machinery on expression level instead of protein level. However, it is not clear how general their results are due to 1) only a handful piRNA genes were identified at that time and 2) they only conducted a test using codon bias but gene expression level was not taken into account. Thus, we performed an analysis of both codon bias and gene expression in all known components of the piRNA machinery and

tested whether the piRNA machinery evolves a higher expression level in *Drosophila* species with higher TE content. Our results show piRNA components evolve both higher gene expression and higher codon bias levels in *Drosophila* species with higher TE load when compared other genes. However, this trend appears to be driven by only a few *Drosophila* lineages. Greater phylogenetic sampling is needed to determine if this signature is robust to phylogenetic non-independence.

Acknowledgements

Here, I would like to express my gratitude to all the people who helped me in completing this thesis. First of all, I would like to extend my deepest gratitude to my P.I., Dr. Justin Blumenstiel, who provided constant guidance and encouragement during my time in his lab. He has walked me through all the stages in my project and this thesis could not have been finished without his help.

I would like to thank Alex and Lucas, who not only gave me help when I confronted difficulties but also made my time in the lab enjoyable. I also would like to thank Michelle and Kendra, who always give me help when I doing my lab work. I would like to thank all my lab mates and people who helped in Haworth. Thanks for let me have an enjoyable time here.

Table of contents

INTRODUCTION

Transposable elements

Transposable elements (TEs) are often referred as "selfish DNA" sequences, which proliferate as a genetic parasite to the detriment of host (Hickey 1982). Although it has been shown that TEs play an important role in evolution, they are still considered to be genome parasites due to their mutagenic characteristics (Kazazian 2004; Bucher, Reinders et al. 2012). This is shown by the fact that certain diseases are caused by TEs' harmful insertion, such as human Factor VIII haemophilia and colon cancer caused by a transposon insertion into the *APC* gene (Kazazian, Wong et al. 1988; Miki, Nishisho et al. 1992).

There are two classes of TE. Class I transposons are also called retrotransposons. They replicate via an RNA intermediate that is reverse transcribed into DNA and finally inserted elsewhere into the genome. Class I transposons are commonly grouped into two main orders: LTRs (long terminal repeats elements) and non-LTRs. Non-LTRs include two subclasses: LINEs (long interspersed elements) and SINEs (short interspersed elements).

Class II transposons are also known as DNA transposons, which use a "cut and paste" mechanism to insert themselves to other positions of the genome. Their transpositions

rely on TE encoded transposase enzymes.

Natural selection against TEs

By having a replicative nature, TEs can proliferate exponentially within the genome across generations (Kazazian 2004). So what controls the exponential proliferation of TEs and what causes variation in TE content across species? For example, 25 to 47% of the genome has been identified as TEs in salamanders and lungfish (Metcalfe and Casane 2013). By contrast, TEs only make up less than 3% of the genome in pufferfish (Aparicio, Chapman et al. 2002). What drives this variation? Furthermore, what prevents TEs from unlimited proliferation? Two major factors likely contribute: 1) natural selection and 2) Small RNA-based genome defense. I will briefly illustrate the recent studies of these two factors below.

Natural selection has been identified as the major factor limiting TEs fixation. Early studies by Brian Charlesworth and colleagues demonstrated that an equilibrium of TEs content could be reached under certain forms of selection on the host. In this model, increasing TE abundance can cause decreasing fitness. The equilibrium in this model is affected by the transposition rate, which determines the increase in TEs copy number and excision rate and natural selection, which determine the decrease in TE copy number. All these factors work together for maintaining the equilibrium number of TEs (Charlesworth and Langley 1986; Charlesworth, Sniegowski et al. 1994). The ectopic recombination model is one of the most reasonable explanations for why increasing TE abundance results in decreased host fitness. A high TE load is harmful to the host's fitness since ectopic exchange among heterozygous TEs can generate tremendously deleterious

chromosomal rearrangements (Montgomery, Huang et al. 1991; Montgomery, Charlesworth et al. 2007). Since ectopic recombination events will increase in frequency with the square of copy number, this provides sufficient synergism to maintain equilibrium copy number. Two other reasonable forces of selection are 1) Deleterious effects of TE expression [TE transcription and translation are very costly(Nuzhdin 1999)] and 2) TE insertional effects through disruption of coding or regulatory regions of the genome.

To what degree does natural selection contribute to variation in TE abundance across species? According to population genetic theory, two major predictions can be made if variation in the strength of natural selection against TE abundance was critical. First, population genetic theory illustrates that genetic drift is weak in species with large population size. Thus, in large population, selection acts more efficiently against the spread of harmful alleles. According to this theory, TE insertions in species with large population size should segregate at very low frequencies. A good example of this case is *Drosophila melanogaster*. Studies have shown that the force of selection against TE much overweighs than strength of genetic drift in *D. melanogaster* (Charlesworth and Langley 1989; Charlesworth, Sniegowski et al. 1994). Since genetic drift becomes very weak in large populations as observed in *D. melanogaster*, the modest variation of population size among related species, in the genus *Drosophila*, might not be important for the TE abundance variation.

Another prediction is that the recombination rate plays an important role in natural selection against TE proliferation. In regions of the genome with low recombination rates, deleterious TE insertions are easy to fix since the lack of recombination reduces the efficacy of natural selection (Hill and Robertso.A 1966).

Piwi-interacting RNAs pathway against TEs

In addition to natural selection, recent studies have shown that mechanisms of TE silencing by small RNAs also play a critical role in limiting TE proliferation (Lippman, May et al. 2003; Aravin, Hannon et al. 2007; Ghildiyal and Zamore 2009). In particular, the Piwi-interacting RNAs (piRNAs) play a dominant role for TE control within animals' germline. The piRNA are a class of small silencing RNAs, typically 24 to 30 nt in length, that form a complex which is found in the Piwi clade of Argonaute (Ago) proteins. They are derived from TEs, and destroy TE transcripts.

In *Drosophila* species, the majority of piRNAs originated from special genomic loci called piRNA clusters, which are often located in pericentromeric heterochromatin. These piRNA clusters store remnant TEs sequences and act as a reservoir of transposons that will be silenced by piRNA machinery (Brennecke, Aravin et al. 2007). Until now, little is known about piRNA clusters. For example, we do not clearly understand how piRNA clusters are transcribed or even how to define piRNA clusters. piRNA clusters can be of two kinds - either dual strand or single strand. For example, in *Drosophila,* only one piRNA cluster is single strand in the somatic follicle cells called *flamenco*. In contrast, most piRNA clusters are specific dual-strand clusters and germline clusters.

In *Drosophila* ovaries, there are two distinct parts of the piRNA pathway: the primary piRNA pathway and the secondary piRNA pathway. In follicle cells, where the somatic cells surround the developing germ cells, only the primary piRNA biogenesis works (Malone, Brennecke et al. 2009). Although the detailed machinery of the piRNA pathway

is not clearly understand, one current model for the primary piRNA pathway is stated as follow: After transcription, the primary piRNA transcripts are shortened into small fragments by the Zucchini endonuclease forming 5' end of piRNA (Nishimasu, Ishizu et al. 2012). Then, these piRNA-like small fragments are loaded onto PIWI family proteins. Recent studies revealed that several piRNA components: Armitage (Armi), Shutdown (Shu) and Vreteno (Vret) are required in the loading process through poorly understood mechanisms (Harris and Macdonald 2001; Handler, Olivieri et al. 2011; Zamparini, Davis et al. 2011; Preall, Czech et al. 2012). In addition to the primary piRNA pathway, an amplification loop works as the secondary piRNA biogenesis in germline. The products of the primary piRNA pathway and maternally deposited piRNAs will trigger the amplification cycle (Ping-Pong) of the secondary pathway. In the Ping-Pong cycle, Aub bound piRNA cleaves the target transposon transcript via slicer activity(Gunawardane, Saito et al. 2007). This process generates the 5' end of new secondary piRNA (Brennecke, Aravin et al. 2007; Brennecke, Malone et al. 2008). The secondary piRNA then loaded on Ago3, and the sense stand of the piRNA-associated Ago3 can match the complementary transposon transcript in the cluster transcript. Transposon transcripts are cleaved, resulting in cytoplasmic transposon silencing, and reused to produce more piRNAs in this amplification cycle.

In *Drosophila,* Piwi and Aub, with their associated primary piRNAs, are maternallly deposited in the embryo and previous study has shown that these maternal deposited piRNAs may play an important role in initiating the ping-pong amplification cycle (Brennecke, Malone et al. 2008). Thus, transgenerational TE control may depend on the dose of maternal piRNA complex that is matched to the TE dose inherited paternally. A good example is provided by the dysgenesis syndrome in *Drosophila virilis*, observed in crosses between strain 9 *Drosophila virilis*, a strain lacking Penelope elements, to strain

160 *Drosophila virilis*, a strain carrying active Penelope elements. Dysgenesis occurs in the offspring from males of strain 160 crossed to strain 9 females. On the other side, the offspring are non-dysgenic from the strain 9 males and strain 160 females. The dysgenic offspring are caused by mobilized Penelope elements. Without maternally inherited piRNAs in offspring stain 9 females, the Penelope elements are out of control. As a result, there is a high probability of gonadal atrophy in offspring from 9 females and 160 males (Blumenstiel and Hartl 2005).

As a newly identified critical factor for TE control, recent studies of the molecular evolution of the piRNA machinery indicate that there is probably a high rate of adaptive evolution between TEs and piRNA machinery, which arises from an arms race between TE invading and host defense, in many *Drosophila* species (Vermaak, Henikoff et al. 2005; Obbard, Gordon et al. 2009; Kolaczkowski, Hupalo et al. 2011). Evolutionary arms races between host and parasite drive adaptions and counter-adaptions against each other, resulting in an increased adaptive evolution rate in host immune systems. In this model, the piRNA machinery functions as an immune system to help the genome guard against TE proliferation within the germline (Vagin, Sigova et al. 2006). Thus, to avoid silencing by the piRNA machinery, TEs might evolve defense strategies against the piRNA silencing. On the other side, if the machinery of genome defense is constantly antagonized by parasites, natural selection will select to counteract these virus' strategies, and this may drive a high rate of adaptive evolution in the protein coding components of the piRNA machinery (Castillo, Mell et al. 2011). However, previous results showed that species with greater TE content have greater levels of purifying selection in the piRNA machinery, measured by ω (the ratio of non-synonymous substitution rates to synonymous substitution rates) (Castillo, Mell et al. 2011). This is the opposite of what might be expected under an evolutionary arms race model. Another important result from

their research is that increasing TE content is correlated with greater codon bias in the piRNA machinery, which is predicted if increasing TE load selects for increased efficiency of host genome defense (Castillo, Mell et al. 2011). Codon bias is associated with increased levels of gene expression (Camiolo, Farina et al. 2012). For example, codon usage bias and gene expression are correlated in fission yeast (Hiraoka, Kawamata et al. 2009). There are only 20 different amino acids but a total of 61 codons encoding them. This means that some amino acids are encoded by more than one codon. These kinds of differences in the frequency of synonymous codons that encode the same amino acids is called codon usage bias (Suzuki, Saito et al. 2009). Thus, it is thought that selection may favor those optimal codons to coadapt with tRNA levels, which can optimize the translation rate and accuracy. Moreover, highly expressed genes experience greater translational selection, associated with the optimal tRNA pool (Bulmer 1991). Thus, we predict that there is an evolutionary arms race between TEs and genome defense provided by the piRNA machinery on expression level instead of protein level. However, Castillo's results are limited because 1) They did not include gene expression level and 2) They only use a handful of piRNA machinery genes in their research. For these reasons, we still do not know how robust their result is.

In our research, we tested if increasing TE content selects for increased function of the piRNA machinery in 12 *Drosophila* species. The effective population size for *D. melanogaster* is large and previous studies have shown that genetic drift may be much weaker than selection in the role of against TEs (Charlesworth and Langley 1989; Charlesworth, Sniegowski et al. 1994). Thus, the TE load variation among members of the *Drosophila* genus may be not strongly affected by the variation in population size. To test if increasing TE content selects for increased efficiency in piRNA machinery in *Drosophila* species, we performed correlation tests for both codon bias and TE content

and gene expression. To evaluate changes in gene expression, we used RNA-seq to measure the gene expression levels in the genomes of *Drosophila* species. RNA-seq is a new technology that uses the capabilities of next-generation sequencing for sequencing transcripts. Compared to two other gene expression methods: qPCR and Microarrays, RNA-seq as a new method offers us advantages. Although qPCR technique is highly sensitive, it is usually used to measure the gene expression for a small number of genes. Since we need genome-wide sequencing in our test, microarrays and RNA-seq are more suitable. Many studies show that there is a high concordance between the Microarrays and RNA-seq results (Bottomly, Walter et al. 2011; Sirbu, Kerr et al. 2012; Zhao, Fung-Leung et al. 2014). But RNA-seq still has advantages when compare to microarrays. RNA-seq avoids using specific probes, which avoids the biases caused by hybridization of microarrays. This offers advantages for RNA-seq in detection of novel transcript and other changes and also can be used to compare gene expression across species. Also RNA-seq offers broader dynamic range and better sensitivity than microarrays, which makes RNA-seq better in performance of measuring low abundance transcripts (Zhao, Fung-Leung et al. 2014). With the measurements of genome-wide codon bias and gene expression values in *Drosophila* species, we designed tests to determine whether increasing TE content selects for increased efficiency in piRNA machinery. This was based on a distribution of correlation coefficients between TE abundance and codon bias values for each gene in the genome with clearly define orthologs. We then compared the distribution of correlation coefficients for piRNA components to the genome wide distribution. We also compared the correlation coefficients distribution of each gene ontology terms [GO terms] to the genome-wide distribution. We further performed this analysis for TE content and gene expression.

Recently, an important resource for researching the piRNA pathway was provided

by an RNAi screen for genes involved in TEs repression (Czech, Preall et al. 2013). In their studies, Hannon and colleagues performed an RNAi knockdown on 8396 genes (This is an unbiased gene set) in *D. melanogaster* and measured the derepression of different TEs caused by gene knockdown. This yielded a ranking of all 8396 genes ordered by decreasing derepression values. In this study, it was found that the knockdown of 74 genes can cause strong derepression of one or more transposons and strikingly these 74 genes included most of the already known piRNA genes (Czech, Preall et al. 2013). The top genes on this knockdown table have strongest effect on TE depression. We determined whether the results from the analysis restricted to genes known to play a role in piRNA biogenesis could be a general property of genes involved in regulation of TEs. By using this knockdown table, we found that the genes that are ranked to have the strongest influence on TE depression are also the ones with the strongest correlation between TE content and codon bias.

MATERIALS AND METHODS

Drosophila species

The 12 *Drosophila* species we used in my project are *D. simulans, D. sechellia, D. melanogaster, D. yakuba, D. erecta, D. ananassae, D.pseudoobscura, D. persimilis, D. willistoni, D. mojoavensis, D. virlis* and *D. grimshawi* [For the gene expression study, we use only 11 species without *D. grimshawi* since *D. grimshawi* strains were not available]. Twelve species coding sequences and genome sequences were downloaded from Flybase (http://flybase.org/). The flies used in RNA-seq study are worldwide samples obtained from the species stock center and kept in room temperature.

Effective number of codons (ENC) estimate

The 12 *Drosophila* species (*D. simulans, D. sechellia, D. melanogaster, D. yakuba, D. erecta, D. ananassae, D.pseudoobscura, D. persimilis, D. willistoni, D. mojoavensis, D. virlis* and *D. grimshawi*) coding sequences were downloaded from Flybase (http://flybase.org/). TE content for each species was measured based on the amount of assembled euchromation that was comprised of repeats estimated from the 12 *Drosophila* species genome consortium (Clark, Eisen et al. 2007; Castillo, Mell et al. 2011). The *D. melanogaster'*s GO term table and orthologs table are also downloaded from Flybase. The effective number of codons (ENC) is a simple measure of how far a gene's codon

usage different from synonymous codons usage (Wright 1990). For all genes, ENC values are ranged from 20 to 61. High ENC value represents a small codon bias value and small ENC value represents a high codon bias value. For example, an ENC value of 20 for a gene sequence means only one codon is used from each synonymous codon group and the has a maximum bias. To contrast, an ENC value of 61 means all codons are contributed equally for encoding and no codon bias. We calculated all genes ENC values in *Drosophila* species by using the CodonW program was written by John Peden. CodonW is a program wrote to simplify Multivariate analysis of codon and amino acid usage (codonw.sourceforge.net). One problem might be arbitrary in this study is that not all genes in *D. melanogaster* have exactly 11 orthologs in other *Drosophlia genus*. First, when the number of orthologs number was less than 12, we remove gene with orthologs number less than seven in our testing since short of such orthologs will highly affect the accuracy of our calculation of the correlation coefficient values and those values could result in an incorrect result for our testing. Secondly, some genes have more than one orthologs in one species, which probably caused by gene fragment or gene duplications. We kept the largest value of codon bias for this situation.

RNA-sequencing

To test if increased TE content can result in selection for increased gene expression of piRNA machinery components, we performed an RNA-sequencing experiment. Eleven of the 12 *Drosophila* species above were used in our RNA-sequencing experiment (Without *D. grimshawi* since we lacked a strain). Each experiment was performed in duplicate. We performed two lanes of sequencing and each lane contains 11 samples from those 11 different *Drosophila* species. The flies are worldwide samples and kept at

room temperature. The RNAs were extracted from 0-2 hours embryos since the piRNAs have been reported to be maternally deposited in the embryo (Akkouche, Grentzinger et al. 2013). The embryos were stored in 1.7ml tubes at -80℃. The concentrations of RNA samples were quantified by using the Qubit method. Each RNA sample contains at least 1 ug RNA to fill the RNA sequencing requirement. All samples were sent to genome sequencing institute and sequenced by Illumina RNA-seq workflow.

Afterward, we used CLC to do the RNA-sequencing analysis: genome annotation and transcript quantification. The data trim and data analysis including the calculation of the correlation coefficients and Mann-Whitney U test were finished by using R programming. Also, not all genes in *D. melanogaster* have exactly 11 orthologs found in other *Drosophila genus*. For gene's ortholog number less than 12, we deleted genes which ortholog number less seven as what we did for ENC testing. For gene's ortholog number larger than 12, which means there are more than one ortholog genes in the same species, we calculated the means to deal with those values. Gene expression is quantified by Read Per Kilobase per Million mapped reads (RPKM).

Bootstrap hypothesis test

Recent studies using a genome-wide screen identified all genes regulating TEs in *Drosophila*. This study provided a list containing 8396 genes ordered by decreasing derepression values of TE (Czech, Preall et al. 2013).

We completed this test by using a bootstrap hypothesis test and the bootstrapping is completed by R programming. The aim was to test whether genes that are ranked as to

having the strongest influence on TE depression were also the ones with the strongest correlation between TE content and codon bias or expression level. In our test, the distribution of correlation coefficients was obtained for each "top bin" that included the top set of genes with the ranked highest effects on TE expression. In other words, the first “top bin” contains the top 10 ranked genes in the knockdown table, the second “top bin” contains the top 11 genes and so forth until the last "top bin" contains the top 200 genes. Overall, 190 "top bins" were obtained. The "top bins" contained gene sets ranked highest for their role in TE repression. For each of these "top bins", we compared the distributions of correlation coefficients between TE and piRNA machinery codon bias and gene expression to 10,000 randomly selected sets equivalent in size to each "top bin" set. A P value was obtained by determining the proportion of times the randomly selected set provided a distribution of correlation coefficients that was greater than the "top bin" set. This approach allowed us to determine whether the set of genes with the strongest effects on TE expression were also ones with significant correlations between codon bias, expression and TE content. By choosing "top bins" of increasing size, we eliminate the use of arbitrary cutoffs for determining the top genes from the knockdown study with an effect on TE expression.

Mann-Whitney U test

To understand if piRNA machinery components evolve a high expression level, on both the codon usage bias level and the gene expression level, in *Drosophila* species with increased TE abundance, we performed a Mann-Whitney U test comparing the Pearson correlation coefficient distributions between the piRNA genes and the whole genome. We first tested if the distribution of correlation coefficient of the piRNA machinery genes

was significantly different from the genome-wide distribution. Then we did this test for all 6300 GO-terms in *D. melanogaster* to detemine if the piRNA GO-term ranked in the top of 6300 GO-terms with a significant positive correlation between 1) TE content and codon bias 2) TE content and gene expression. The Mann-Whitney U test was performed using R programming.

RESULTS

The piRNA machinery displays a higher codon bias in genomes with greater TE content

To understand whether increasing TE contents selects for increased efficiency in piRNA machinery, we performed a correlation-based test between TE abundance and codon bias for genes in 12 *Drosophila* species. At first, on a genome wide level, we calculated codon bias value (ENC) for all the genes from 12 *Drosophila* species. For each gene in *D. melanogaster* and its orthologs in the other 11 species, we calculated the Pearson correlation coefficient between ENC and TE content. Also, from recent studies, we obtained a list containing 26 already known piRNA genes (Supplemental table 1). For these piRNA genes, we found that 15 of 26 genes show a negative correlation between ENC and TE abundance (larger ENC represents weaker codon bias). More strikingly, we found that 10 piRNA genes on the list are in the range of top 1000 of total 11237 genes (Supplemental table 2). These results show that codon bias and TE load have a negative correlation on the genome average level, but the majority of the piRNA components show a positive correlation between codon bias and TE content (Figure 1A). On whole genome level, we find that only 3915 of 11237 genes show a positive correlation between codon bias and TE content (larger ENC value represents smaller codon bias value) and the average correlation coefficient for these 11237 is 0.127509 (Figure 1B). This result shows that on the genome average level, codon bias and TE abundance show a slightly negative correlation. To test the evolutionary arms race model, the piRNA machinery components were compared to the rest of the genome with respect to increased codon

bias in species with high TE content. Thus, the Pearson correlation coefficient distribution of piRNA pathway components should be significantly different than the distribution for the whole genome. We performed a Mann-Whitney U test for comparing the distribution of correlation coefficients between the piRNA pathway genes and the distribution of whole genome (Figure 1). The result shows that the distribution for the correlation coefficients of the piRNA genes is significantly different from the distribution of correlation coefficients for genome wide (p-value = 0.0074587). While piRNA machinery components might be different than background, it is also critical to compare this functional class to all other functional classes. To test if the piRNA machinery components are a unique category of gene classes with respect to this trend, an approach using gene ontology terms (GO terms) was implemented. GO terms is an ontology of defined terms that representing gene properties. GO terms are specified in three categories: cellular component, molecular function and biological process. By using the Mann-Whitney U test, we compared the correlation coefficients distribution of every GO-term in *D. melanogaster* against the background distribution of correlation coefficients. By sorting the P values, we found that the piRNA GO-term is ranked in top 17 of 6300 GO terms with a significant positive correlation between TE content and codon bias (Figure 2). This result supports our hypothesis that piRNA components evolve a higher codon bias level in higher TE load *Drosophila* species.

Recently, an important resourse for researching the piRNA pathway was provided by an RNAi screen for genes involved in TE expression(Czech, Preall et al. 2013). In this analysis, they performed germline RNAi knockdown on 8396 genes in *D. melanogaster* and measured the derepression of different TEs. They found that the knockdown of 74 genes could cause strong derepression of one or more transposons and most piRNA genes

are ranked in the top 100, especially top 50, of this list of genes sorted by their strength of TE repression (Supplemental table 3) (Czech, Preall et al. 2013). To determine whether the observed increase in codon bias was a general property of all genes regulating TEs rather than simply genes involved in piRNA biogenesis we extended our analysis to the gene set ranked by strength of TE derepression during knockdown. We performed a bootstrapping test for testing to determine whether genes that are ranked as to have the strongest influence on TE depression are also the ones with the strongest correlation between TEs and codon bias. To perform this test, we selected different size bins (of increasing size) of the genes that were top ranked in their ability to repress TEs. For example, a set of the top 10 genes, the top 11 genes, the top 12 genes and so forth were selected. For each of these top ranked gene sets, we performed a bootstrap test of significance by randomly sampling 10,000 gene sets of similar size, and comparing the proportion of times the randomly selected sets showed a stronger correlation between TE content and codon bias. This served to provide an empirical P-value for a test of significance for the relationship between TE content and codon bias for each of the top gene sets. The bootstrapping test result shows that the top genes in the knockdown table, which can cause strong derepression in Hannon's experiment, also have significantly greater correlation between TE content and ENC (Figure 3).

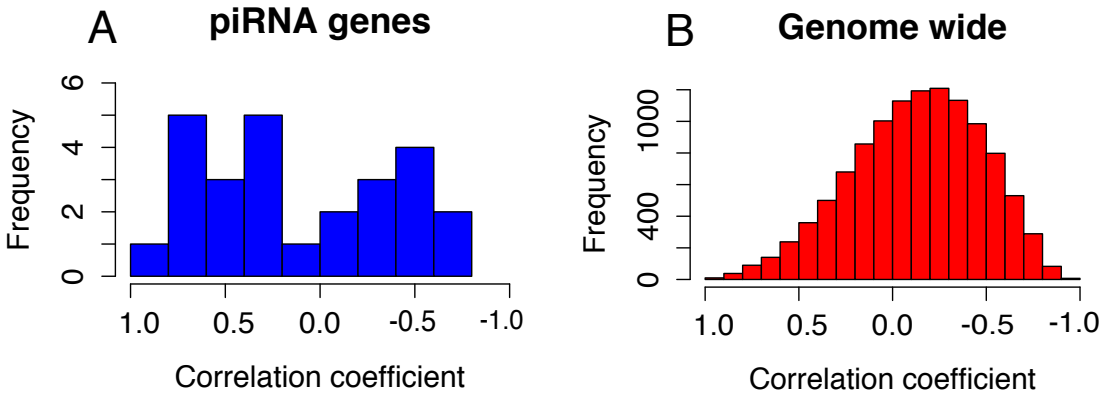

Figure 1: **The distribution of correlation coefficients between codon bias and TE abundance**. A: The distribution of correlation coefficients for piRNA genes between codon bias and TE load. B: The distribution of the whole genome.

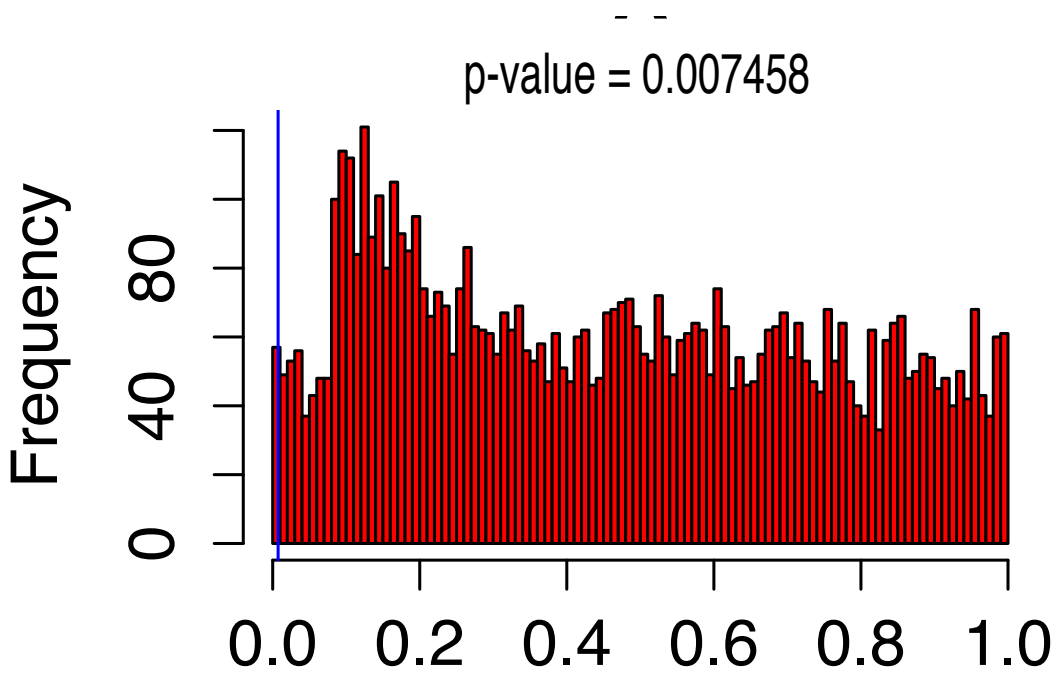

Figure 2: **The distribution of the p-value of GO-terms.** The p-values are the Mann-Whitney U test results for all GO-terms**.** Each p-value represents if, for the related GO-term, the distribution of the Pearson correlation coefficient between ENC and TE load significantly different to the distribution of the whole genome. The blue line marks indicates the location of the p-value of the piRNA GO-term.

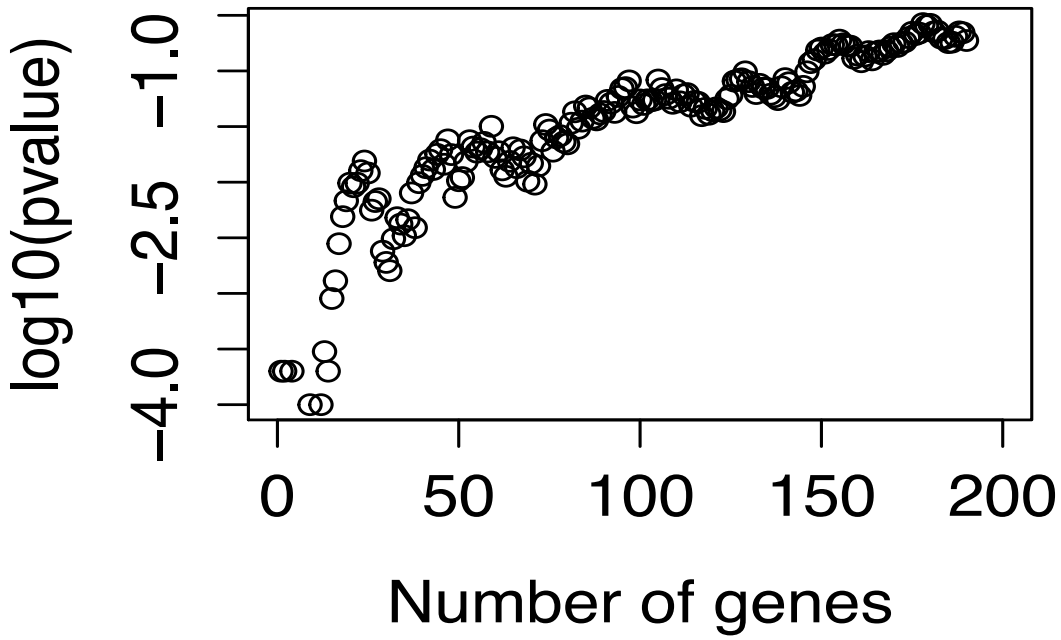

Figure 3**: The bootstrap hypothesis results using the ENC calculated Pearson correlation coefficients.** The p-values are log10(p-value). Each cycle in the figure is a "top bin" set we used in our bootstrap hypothesis test. The value on the x-axis is how many top genes (on the knockdown table) in the related "top bin" set. In our test, the p-value on the y-axis was obtained by determining the proportion of times the random selected set provided a distribution of correlation coefficients that greater than the "top bin" set (details in method). The test results support our hypothesis that the top lists on the knockdown table with strongest influence on TE depression also have relative the strongest correlation coefficients between codon bias and TE content. The significance goes away with the increased size of "top bin" set.

The piRNA machinery components evolve higher gene expression with increasing TE

To understand whether the piRNA machinery displays higher gene expression under higher TE burden, we also performed RNA-sequencing for 0-2 hour old embryos to quantify gene expression for 11 *Drosophila* species. The gene expression is represented by reads per kilobase per million reads (RPKM) in this paper. We calculated Pearson correlation coefficient between RPKM and TE load for *D. melanogaster* and its orthologs. For the piRNA components, 14 of 26 genes show positive correlation between RPKM and TE abundance and the mean correlation coefficient of these 26 genes equals 0.03416605 (Figure 4A). By contrast, we found that the distribution of correlation coefficients for whole genome is slightly biased to the negative side (Figure 4B). There are 3958 of 11255 genes that show a positive correlation between RPKM and TE content, and the mean correlation coefficient is -0.057. There was a significant difference in the distribution of correlation coefficients between TE burden and gene expression for the piRNA components compared to the distribution of correlation coefficients for all genes ($p = 0.047$). To test for significance, we performed a Mann-Whitney U test for all GO-terms in *D. melanogaster*. By sorting the p-values from our test, the piRNA GO term was also found enriched in 6300 GO terms on the positive correlation side (Figure 5). This result shows that piRNA components also evolve a higher gene expression level with higher TE abundance.

To determine whether our observed increase in gene expression was a general property of all genes regulating TE rather than simply the piRNA biogenesis components, we tested whether those genes on the knockdown table ranked as having the strongest influence on TE depression also the ones with strongest correlation between TE content and gene expression (Supplemental table 4). However, the bootstrap hypothesis test shows that those genes, on the top list of the knockdown table, do not have significantly greater Pearson correlation coefficient between TE and RPKM (Figure 6).

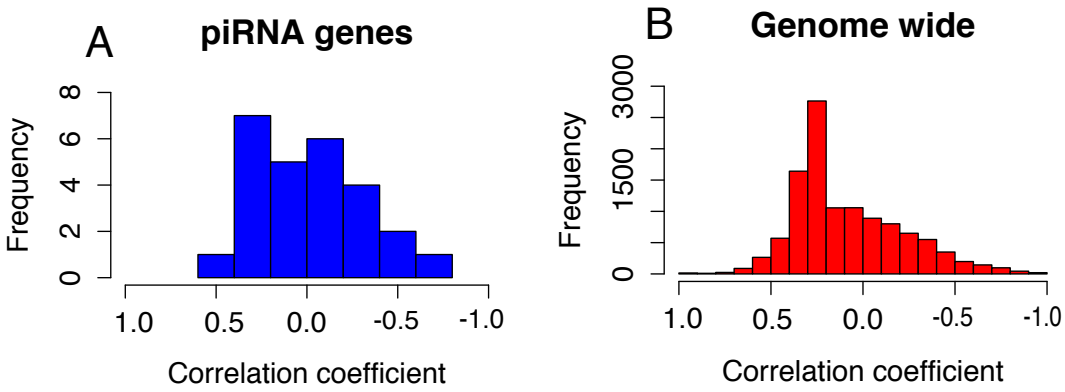

Figure 4: **The distribution of correlation coefficients between TE content and RPKM.** A: The distribution of correlation coefficients for piRNA genes between RPKM and TE load. B: The distribution of correlation coefficients between RPKM and TE load for genes in whole genome wide

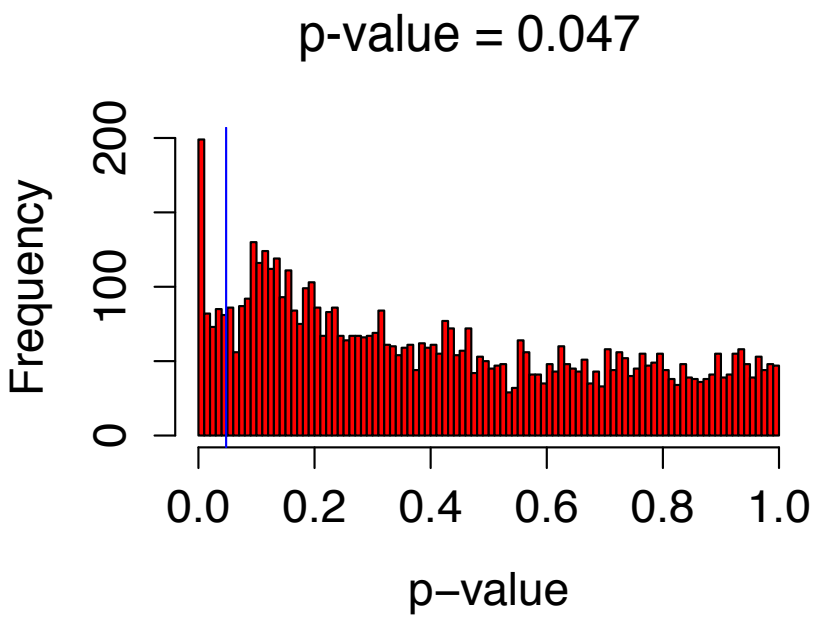

Figure 5: **The distribution of GO-term p-value, which calculated by using the correlation coefficient between RPKM and TE load.** The p-values are the Mann-Whitney U test results for all GO-terms**.** Each p-value represents if, for the related GO-term, the distribution of the Pearson correlation coefficient between RPKM and TE load significant different to the distribution of the whole genome. The blue line marks where is the piRNA GO-term's p-value.

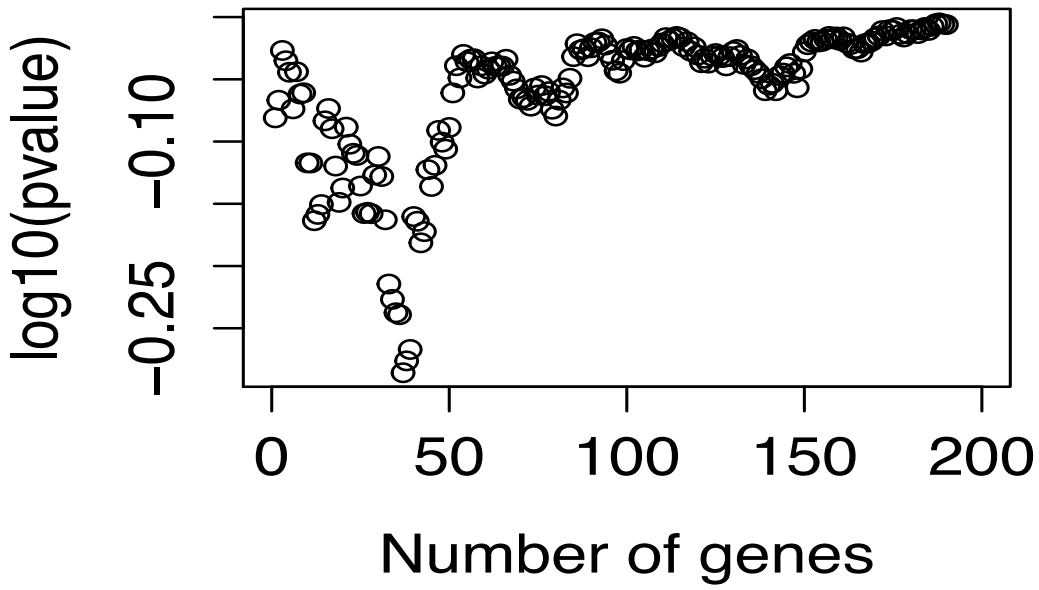

Figure 6: **The bootstrap hypothesis test results by using the RPKM calculated correlation coefficient**. The The p-values are log10(p-value). Each cycle in the figure is a "top bin" set we used in our bootstrap hypothesis test. The value on the x-axis is how many top genes (on the knockdown table) in the related "top bin" set. In our test, the p-value on the y-axis represents the proportion of times the random selected set provided a distribution of correlation coefficients that greater than the "top bin" set (details in method).

Phylogenetic independent contrast results

Closely related species will tend to show similarities in traits, this tendency is phylogenetic signal. Thus, the value of traits might not independent. Phylogenetic independent contrast (PIC) is a method uses phylogenetic information for transforming interspecific data into independent and identically distributed values.

To determine whether results were robust to phylogenetic non-independence, we performed phylogenetic independent contrasts by the R package "APE". We calculated the PICs for ENC, RPKM and TE content. However, the distribution of correlation coefficients for piRNA components was not significantly different compared to the genome distribution on both codon bias level and gene expression level ($p = 0.37$ and $p = 0.41$). One possible explanation is that the correlation coefficients before phylogenetic correction might be strongly affected by some long branches with high TE content. After removing the data of *D. ananassae* and *D. willistoni,* two species with high TE content and long branches, we recalculated the ENC-TE contrast correlation coefficient again. After comparing the new distribution of correlation coefficients to previous values, we found that 7 of 10 piRNA components are strongly affected by the two long branches. Also, the remaining three also lost high correlation coefficients after removing the *D. willistoni* branch. Thus, much of the results is explained by these two high TE abundance lineages and phylogenetic correction reduce the power of the test. Thus, in a comparison to genes across the genome, the piRNA machinery shows a signature of increased efficiency in species with high TE content. However, this trend appears to be driven by a contrast of one or two lineages rather than across the entire genus.

Discussion

TEs are considered "selfish DNA" sequences, which act as parasites to the detriment of host (Hickey 1982). Thus, the co-evolutionary dynamic between TEs and their hosts are interesting for scientists. Some interesting questions are: What causes the TE variation across species? What prevents TEs from unlimited proliferation? Why is the impact of TEs great in some species' genomes but some other species' genomes are minimally impacted? First of all, population genetic theory has indicated that natural selection is one major factor contributing to TE control. To what degree does the strength of natural selection contribute to the TE variation in species? According to population genetic theory, genetic drift is weak with large population size. Previous studies have proved that the force of selection against TE greatly overweighs the effect of drift in *D. melanogaster* (Charlesworth and Langley 1989; Charlesworth, Sniegowski et al. 1994). Thus, in species with large population size like *Drosophila* species, modest variation of population size among related species is probably not an important reason for TE variation across species. As a result, in the *Drosophila* genus, the piRNA machinery as a genome defense pathway may play the most important role for TE regulation.

Recent studies of the molecular evolution of the piRNA machinery indicate that there is probably a high rate of adaptive evolution between TEs and piRNA machinery, which arises from an arms race between TE invading and host defense, in many *Drosophila* species (Vermaak, Henikoff et al. 2005; Obbard, Gordon et al. 2009; Kolaczkowski, Hupalo et al. 2011). Thus, we predicted that there is an evolutionary arms race between TE content and genome defense provided by the piRNA pathway. However, previous

work has shown that within the *Drosophila* genus species with greater TE load have greater levels of purifying selection, measured by ω (the ratio of non-synonymous substitution rates to synonymous substitution rates) in the piRNA machinery (Castillo, Mell et al. 2011). This is the opposite of what might be expected under the evolutionary arms race model. Furthermore, increasing TE content was found correlated with greater codon bias in the piRNA machinery in *Drosophila species*, which is predicted if increasing TE load selects for increased efficiency of host genome defense (Castillo, Mell et al. 2011). This result may indicate that there is an evolutionary arms race between TEs and the piRNA silencing pathway in expression level instead of protein level. However, their results were limited for two reasons: 1) they only used a handful piRNA genes in their test. 2) They only tested codon bias level but ignored gene expression level. To understand the co-evolutionary dynamics between TE and piRNA machinery and test how general their result is, we completed our test with an updated piRNA genes list and added a test of gene expression.

For codon usage bias, we conclude that in 12 *Drosophila* species, the distribution of correlation coefficients between TE content and codon bias in the piRNA genes is significantly different compared to the distribution in the whole genome. Furthermore, piRNA GO term is significantly different relative to the rest of the GO terms with respect to increased codon bias in species with higher TE content. This results show that the piRNA pathway components evolve a higher codon bias expression level with higher TE abundance in *Drosophila* genus. This means that increasing TE content has perhaps selected for increased expression of piRNA pathway component genes. Moreover, this result supports our evolutionary arms race model that under relatively large population sizes, as seen in the *Drosophilia* genus, the dynamic of TEs and the host defense

provided by the piRNA pathway can be explained by an evolutionary arms race. Evolutionary arms races between host and TEs drive adaptions and counter-adaptions against each other, resulting in an increased adaptive evolution rate in the piRNA silencing pathway. On the side of TEs, there may be strong selection to antagonize the piRNA silencing pathway by evolving defense strategies for avoid silencing. On the other side, under the antagonized strategies by TEs, natural selection acts on counteracting these strategies to protect the host, which will drive a high adaptive evolution rate for piRNA machinery components.

Furthermore, we can conclude that the piRNA pathway components evolve a higher gene expression level with higher TE abundance in 11 *Drosophila* species. In our test, although the signal of the result is not as strong as we got from the codon bias test, the distribution of the correlation coefficient between TE content and gene expression in the piRNA pathway components significantly differ from the distribution in whole genome and the piRNA GO-term is enriched in more than 6300 GO terms ranked by p-values. Considering these results and the codon bias level's result together, we can conclude that increasing TE content has selected for increased expression of piRNA pathway component genes on both codon usage bias level and gene expression level. And the results from both the codon bias level and the gene expression level can be explained by evolutionary arms race model.

Recently, an unbiased, genome-wide RNAi screen for genes involved in TE control in *Drosophila* was provided by the Hannon lab (Czech, Preall et al. 2013). Most piRNA genes are in the genes have strongest TE derepression after knockdown. To determine

whether the observed increase in both codon bias and gene expression were a general property of all genes regulating TEs rather than simply genes involved in piRNA biogenesis we extended our analysis to the gene set ranked by strength of TE derepression during knockdown. Our bootstrapping test results show that the genes have the strongest correlation coefficients between TE content and codon bias are the ones also have the strongest influence on TE repression. However, this significant result went away for the similar test on the gene expression level. The reason for this is complicated and unclear. One explanation reason might be that the evolutionary arms race can not explain all evolutionary dynamic between host and parasite (Castillo, Mell et al. 2011). For example, an alternative model is trench warfare. In this model, a diversity of parasite alleles may be selected to maintain some polymorphisms of defense strategies and favor a adaptive fixation (Stahl, Dwyer et al. 1999). Moreover, the correlation coefficients in our table only considered the dynamic between TE and genome defense. However, many genes ranked highly from the RNAi screen are probably involved in other pathways other than TEs and the genome defense, which may affect how selection acts on the expression of genes.

Phylogenetic signal is a tendency that closely related species show similarities in traits. Thus, our correlation coefficient result may not be independent of the effect of phylogenetic signal. To eliminate phylogenetic signal, we calculated phylogenetic independent contrasts and our results were not robust to this phylogenetic correction. However, we found that our phylogenetic correction approach is limited due to the large number of traits being sampled. In our result, each gene represents a trait and thus we have more than 12000 traits. Thus, it is very difficult to identify a trait transformation that is uniform across traits. Future works are needed to develop methods for comparing the

evolution of thousands of traits (ie, gene expression) on a single tree. Importantly, we find that much of our correlation coefficients results are explained by two species with high TE abundance, *D. ananassae* and *D. willistoni*. Thus, phylogenetic correction reduces the power of the test. But, since all genes in the genome share same phylogenetic history, this result shows that on the several lineages with high TE content show a concerted change in the piRNA machinery.

In conclusion, our tests have shown that piRNA pathway components evolve a higher expression level on both codon bias level and gene expression with higher TE abundance. Thus, in species with large population size like *Drosophila,* piRNA silencing pathway may majorly contribute to the TE variation across species. This result supports the evolutionary arms race model as an explanation of the dynamic between TE and genome defense provided by piRNA silencing pathway. Moreover, we provided a table for genes regulating TE content on both codon bias level and gene expression level in *Drosophila* species. This table may provide important information on studying the co-evolutionary dynamic between genome defense and TE in the future study.

References:

Akkouche, A., T. Grentzinger, et al. (2013). "Maternally deposited germline piRNAs silence the tirant retrotransposon in somatic cells." EMBO Reports **14**(5): 458-464.

Aparicio, S., J. Chapman, et al. (2002). "Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*." Science **297**(5585): 1301-1310.

Aravin, A. A., G. J. Hannon, et al. (2007). "The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race." Science **318**(5851): 761-764.

Blumenstiel, J. P. and D. L. Hartl (2005). "Evidence for maternally transmitted small interfering RNA in the repression of transposition in *Drosophila virilis*." Proceedings of the National Academy of Sciences of the United States of America **102**(44): 15965-15970.

Bottomly, D., N. A. R. Walter, et al. (2011). "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays." PloS One **6**(3).

Brennecke, J., A. A. Aravin, et al. (2007). "Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila." Cell **128**(6): 1089-1103.

Brennecke, J., C. D. Malone, et al. (2008). "An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing." Science **322**(5906): 1387-1392.

Bucher, E., J. Reinders, et al. (2012). "Epigenetic control of transposon transcription and mobility in Arabidopsis." Current Opinion in Plant Biology **15**(5): 503-510.

Bulmer, M. (1991). "The Selection-Mutation-Drift theory of synonymous codon usage." Genetics **129**(3): 897-907.

Camiolo, S., L. Farina, et al. (2012). "The Relation of Codon Bias to Tissue-Specific Gene Expression in Arabidopsis thaliana." Genetics **192**(2): 641-+.

Castillo, D. M., J. C. Mell, et al. (2011). "Molecular evolution under increasing transposable element burden in Drosophila: A speed limit on the evolutionary arms race." Bmc Evolutionary Biology **11**.

Charlesworth, B. and C. H. Langley (1986). "The Evolution of Self-Regulated Transposition of Transposable Elements." Genetics **112**(2): 359-383.

Charlesworth, B. and C. H. Langley (1989). "The Population Genetics of *Drosophila* Transposable Elements." Annual Review of Genetics **23**: 251-287.

Charlesworth, B., P. Sniegowski, et al. (1994). "The Evolutionary Dyamics of Repetitive DNA in Eukaryotes." Nature **371**(6494): 215-220.

Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the *Drosophila* phylogeny." Nature **450**(7167): 203-218.

Czech, B., J. B. Preall, et al. (2013). "A Transcriptome-wide RNAi Screen in the Drosophila Ovary Reveals Factors of the Germline piRNA Pathway." Molecular Cell **50**(5): 749-761.

Ghildiyal, M. and P. D. Zamore (2009). "Small silencing RNAs: an expanding universe." Nature Reviews Genetics **10**(2): 94-108.

Gunawardane, L. S., K. Saito, et al. (2007). "A slicer-mediated mechanism for repeat-associated siRNA 5 ' end formation in Drosophila." Science **315**(5818): 1587-1590.

Handler, D., D. Olivieri, et al. (2011). "A systematic analysis of Drosophila TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors." Embo Journal **30**(19): 3977-3993.

Harris, A. N. and P. M. Macdonald (2001). "aubergine encodes a Drosophila polar granule component required for pole cell formation and related to eIF2C." Development **128**(14): 2823-2832.

Hickey, D. A. (1982). "Selfish DNA - A Sexually-Transmitted Nuclear Parasite." Genetics **101**(3-4): 519-531.

Hill, W. G. and Robertso.A (1966). "Effect of Linkage on Limits to Artifical Selection" Genetical Research **8**(3): 269-&.

Hiraoka, Y., K. Kawamata, et al. (2009). "Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe." Genes to Cells **14**(4): 499-509.

Kazazian, H. H. (2004). "Mobile elements: Drivers of genome evolution." Science **303**(5664): 1626-1632.

Kazazian, H. H., C. Wong, et al. (1988). "Hemophilia-A Resulting from *de novo* Insertion of L1 Sequences Represents a Novel Mechanism for Mutation in Man." Nature **332**(6160): 164-166.

Kolaczkowski, B., D. N. Hupalo, et al. (2011). "Recurrent Adaptation in RNA Interference Genes Across the Drosophila Phylogeny." Molecular Biology and Evolution **28**(2): 1033-1042.

Lippman, Z., B. May, et al. (2003). "Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification." Plos Biology **1**(3): 420-428.

Malone, C. D., J. Brennecke, et al. (2009). "Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary." Cell **137**(3): 522-535.

Metcalfe, C. J. and D. Casane (2013). "Accommodating the load: The transposable element content of very large genomes." Mob Genet Elements **3**(2): e24775.

Miki, Y., I. Nishisho, et al. (1992). "Disruption of the APC Gene by a Retrotransposal Insertion of L1 Sequence in a Colon Cancer." Cancer Research **52**(3):

643-645.

Montgomery, E., B. Charlesworth, et al. (2007). "A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster (Reprinted)." Genetics Research **89**(5-6): 435-445.

Montgomery, E. A., S. M. Huang, et al. (1991). "Chromosome Rearrangement by ectopic recombination in *Drosophila melanogaster* - Genome Structure and Evolution." Genetics **129**(4): 1085-1098.

Nishimasu, H., H. Ishizu, et al. (2012). "Structure and function of Zucchini endoribonuclease in piRNA biogenesis." Nature **491**(7423): 284-U157.

Nuzhdin, S. V. (1999). "Sure facts, speculations, and open questions about the evolution of transposable element copy number." Genetica **107**(1-3): 129-137.

Obbard, D. J., K. H. J. Gordon, et al. (2009). "The evolution of RNAi as a defence against viruses and transposable elements." Philosophical Transactions of the Royal Society B-Biological Sciences **364**(1513): 99-115.

Preall, J. B., B. Czech, et al. (2012). "shutdown is a component of the Drosophila piRNA biogenesis machinery." Rna-a Publication of the Rna Society **18**(8): 1446-1457.

Sirbu, A., G. Kerr, et al. (2012). "RNA-Seq vs Dual- and Single-Channel Microarray Data: Sensitivity Analysis for Differential Expression and Clustering." Plos One **7**(12).

Stahl, E. A., G. Dwyer, et al. (1999). "Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis." Nature **400**(6745): 667-671.

Suzuki, H., R. Saito, et al. (2009). "Measure of synonymous codon usage diversity among genes in bacteria." Bmc Bioinformatics **10**.

Vagin, V. V., A. Sigova, et al. (2006). "A distinct small RNA pathway silences selfish genetic elements in the germline." Science **313**(5785): 320-324.

Vermaak, D., S. Henikoff, et al. (2005). "Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila." Plos Genetics **1**(1): 96-108.

Wright, F. (1990). "The 'effective number of codons' used in a gene." Gene **87**(1): 23-9.

Zamparini, A. L., M. Y. Davis, et al. (2011). "Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in Drosophila." Development **138**(18): 4039-4050.

Zhao, S. R., W. P. Fung-Leung, et al. (2014). "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells." Plos One **9**(1).